Global Journal of Computer Science and Technology
Volume 12 Issue 9 Version 1.0 April 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Advanced Methods to Improve Performance of K-Means Algorithm: A Review

By Ritu Yadav & Anuradha Sharma
Guru Jambheshwar University of Science & Technology,Hissar,Haryana India

Abstract - Clustering is an unsupervised classification that is the partitioning of a data set in a set of meaningful subsets. Each object in dataset shares some common property- often proximity according to some defined distance measure. Among various types of clustering techniques, K-Means is one of the most popular algorithms. The objective of K-means algorithm is to make the distances of objects in the same cluster as small as possible. Algorithms, systems and frameworks that address clustering challenges have been more elaborated over the past years. In this review paper, we present the K-Means algorithm and its improved techniques.

Keywords : *classification, clustering, k-means clustering, partitioning clustering.*

GJCST Classification: *F.2,F.2.m*

Strictly as per the compliance and regulations of:

Advanced Methods to Improve Performance of K-Means Algorithm: A Review

Ritu Yadav[α] & Anuradha Sharma[σ]

Abstract - Clustering is an unsupervised classification that is the partitioning of a data set in a set of meaningful subsets. Each object in dataset shares some common property- often proximity according to some defined distance measure. Among various types of clustering techniques, K-Means is one of the most popular algorithms. The objective of K-means algorithm is to make the distances of objects in the same cluster as small as possible. Algorithms, systems and frameworks that address clustering challenges have been more elaborated over the past years. In this review paper, we present the K-Means algorithm and its improved techniques.

Keywords : classification, clustering, k-means clustering, partitioning clustering.

I. Introduction

Clustering is a type of categorization imposed rules on a group of data points or objects. A broad definition of clustering could be "the process of categorizing a finite number of data points into groups where all members in the group are similar in some manner". As a result, a cluster is a aggregation of objects. All data points in the same cluster have common properties (e.g. distance) which are different to the data points laying in other clusters.

Cluster analysis is an iterated process of knowledge discovery and it is a a multivariate statistical technique which identifies groupings of the data objects based on the inter-object similarities computed by a chosen distance metric .Clustering algorithms can be classified into two categories: Hierarchical clustering and Partitional clustering [1]. The partitional clustering algorithms, which differ from the hierarchical clustering algorithms, are usually to create some sets of clusters at start and partition the data into similar groups after each iteration. Partitional clustering is more used than hierarchical clustering because the dataset can be divided into more than two subgroups in a single step but for hierarchy method, always merge or divide into 2 subgroups, and don't need to complete the dendrogram[2].

Cluster analysis of data is an important task in knowledge discovery and data mining. Cluster analysis aims to group data on the basis of similarities and dissimilarities among the data elements. The process can be performed in a supervised, semi-supervised or unsupervised manner. Different algorithms have been proposed which take into account the nature of the data and the input parameters in order to partition the data. Data vectors are clustered around centroid vectors. The cluster the data vector belongs to is determined by its distance to the centroid vector. Depending on the nature of the algorithm, the numbers of centroids are either defined in advance by the user or automatically determined by the algorithm. Discovering the optimum number of clusters or natural groups in the data is not a trivial task. The popular clustering techniques which are suggested so far are either partition based or hierarchy based, but both approaches have their own advantages and limitations in terms of the number of clusters, shape of clusters, and cluster overlapping[3] .Some other approaches are designed using different clustering techniques and involve optimization in the process. The involvement of intelligent optimization techniques has been found effective to enhance the complex, real time, and costly data mining process.

II. K-Means Algorithm

The conventional K-mean algorithm is based on decomposition, most popular technique in data mining field. The concept of K-Means algorithm uses K as a parameter, Divide n object into K clusters, to create relatively high similarity in the cluster and, relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The measurement of the similarity for the algorithm selection is done by the reciprocal of Euclidean distance. That is to say, the closer the distance, the bigger the similarity of two objects, and vice versa.

a) Procedure of K-mean Algorithm

K-mean distributes all objects to K number of clusters at random;

1) Calculate the mean value of each cluster, and use this mean value to represent the cluster;
2) Re-distribute the objects to the closest cluster according to its distance to the cluster center;

Author α : Ritu Yadav , Department of Computer Science and Engg,Guru Jambheshwar University of Science & Technology,Hissar,Haryana India. E-mail:ryadav1986@gmail.com
Author σ : Anuradha Sharma,aDepartment of Computer Science and Engineering , Rajasthan Institute of Engineering & Technology, Rajasthan Technical University,Jaipur. Rajasthan,India. E-mail:anuradha.sharma918@gmail.com

3) Update the mean value of the cluster, say, calculate the mean value of the objects in each cluster;
4) Calculate the criterion function E, until the criterion function converges.

Usually, the K-mean algorithm criterion function adopts square error criterion, defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

In which, E is total square error of all the objects in the data cluster, p is given data object, mi is mean value of cluster Ci (p and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible [4].

b) Analysis of the Performance of the K-mean Algorithm

i. *Advantages*

1) It is a classic algorithm to resolve cluster problems; this algorithm is simple and fast;
2) For large data collection, this algorithm is relatively flexible and highly efficient, because the complexity is O (ntk), among which, n is the number of all objects, k is the number of cluster, t is the times of Iteration. Usually, k<<n and t<<n. The algorithm usually ends with local optimum.
3) It provides relatively good result for convex cluster;
4) Because of the limitation of the Euclidean distance [5]. It can only process the numerical value, with good geometrical and statistic meaning;

ii. *Disadvantages*

1) Sensitive to the selection of initial cluster center, usually end without global optimal solution, but suboptimal solution;
2) There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated;
3) This algorithm is easy to be disturbed by abnormal points; a few of this abnormal data will cause extreme influence to the mean value;
4) Sometimes the result of cluster may lose balance

III. Advanced K-Means Clustering Algorithms

The K-means algorithm and its conjoined algorithms are in the family of center base clustering algorithms. This family have several methods: expectation maximization, fuzzy K-means and harmonic K-means. A brief review of these algorithms is given in the following sub-sections.

a) General Clustering Algorithm

D. Mayszko [6], proposed the steps for the clustering algorithm are as follows:

1) Initialize step with centers C.
2) For each data point xi, compute its minimum distance with each center cj.
3) For each center cj, recomputed the new center from all data points xi belong to this cluster.
4) Repeat steps 2 and 3 until convergence.

b) Expectation Maximization

According to C.M. Bishop, Expectation maximization algorithm uses a linear combination of Gaussian distribution as centers [7]. Its minimization is:

$$EM(X,C) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} p(x_i|c_j)p(c_j)\right)$$

This algorithm has a constant weight that gives all data point to its nearest center.

c) Fuzzy K-Means

According to S. Wierzchoń, Fuzzy K-means algorithm is also called fuzzy c-means. It is adaptation of the K-means algorithm and use soft membership function. This algorithm determines a data point belongs to any centers depends on its membership as [6]:

$$FKM(X,C) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \|x_i - c_j\|^2$$

This algorithm has a soft membership and constant weight that gives all data point to the closed center.

d) Harmonic K-Means Algorithm

According to B. Zhang, The harmonic K-means algorithm is a method which is similar to the standard K-means. It uses the harmonic mean of the distance from each data point to all centers as [8]:

$$HKM(X,C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^2}}$$

This algorithm has a soft membership and weight function to points that are far away from every center.

e) Early Stop K-Means

Early stop K-means algorithm is the first one to handle a convergence step in the standard K-means algorithm. It consists of associating the square error values to a convergence condition. It gets action when there are two consecutive iterations and the square error of the last iteration exceeds that of the preceding iteration. It finds a solution at least as good as that of the standard K-means with a number of iterations smaller than or equal to that of standard K-means algorithm [9].

f) Modified K-Means

According to W. Li, Modified K-means algorithm is a new algorithm for K-means based on the

optimization formulation and a novel iterative method. The steps of this algorithm represented as [10]:

1) Dividing data set (D) into K parts:

$$D = \cup_{k=1}^{K} S_k, S_{k1} \cap S_{k2} = \emptyset, k1 \neq k2$$

2) Let $x_{(0)}^{(0)}, k = 1, \ldots, K$ be initial clustering centers calculate by:

$$x_{(0)}^{(0)} = \sum_{d^{(j)} \in S_k} d^{(j)} / |S_k|, \quad k = 1, \ldots, K$$

3) Decide membership of the patterns in each one of the K clusters according to the minimum distance from cluster center.
4) Calculate new centers using the iteratively.
5) Repeat step 3 and 4 till there in no change in cluster center.

IV. Methods to Improve K-Means Algorithm's Performance

a) *Methods for Initial Point Selection*

i. *Refining Initial Points Algorithm*

In partitional clustering algorithm, the first step we should get initial seed points (cluster centers). To choose good initial points will improve solutions and reduce execution time. Refining initial points algorithm is proposed[2].

For a start, randomly choose some subsets within equal number of samples from large data sets. Secondly, partitional algorithm is applied to each subsets to get each center sets of the subsets. Thirdly, gather these center sets and apply the partitional algorithm again to obtain the most proper center set. For getting good initial seed points, totally we repeat the partitional algorithm 2 times by fewer sample sets. Finally run the partitional algorithm with the most feasible center set as the initial seeds and original large data sets.

The algorithm steps are:

1) Randomly build J sample subsets. Si is a random subset of data (i = 1...J and the size of Si is S_size).
2) Use modified algorithm to find center Ci of each Si . Gather all Ci (i = 1...J) into C _Total .
3) For each set Ci (i = 1...J), run paritional algorithm with initial points Ci and data set C _Total to get another center set FCi .
4) For each FCi(i = 1...J), calculate sum Sumi of the distance between each point in C_Total to the closest center point in FCi .
5) Find minimum of Sumi (i =1...J). If Sump is minimum, take FCp as final initial points.

ii. *Cluster Centroid Decision Method*

This method proposed a technique to assign the data point to appropriate cluster's centroid, we calculate the distance between each cluster's centroid and for each centroid take the minimum distance from the remaining centroid and make it half, denoted by dc(i) i.e. half of the minimum distance from ith cluster's centroid to the remaining cluster's centroid. Now take any data point to calculate its distance from ith centroid and compare it with dc(i). If it is less than or equal to dc(i) then data point is assigned to the ith cluster otherwise calculate the distance from the other centroid. Repeat this process until that data point is assigned to any of the remaining cluster. If data point is not assigned to any of the cluster then the centroid which shows the minimum distance with data point becomes the cluster for that data point. Repeat this process for each data point. Take mean of each cluster separately and update the centroid of clusters like traditional k-mean. Repeat this process until termination condition is achieved[3].

N0: Number of data point
K : Number of cluster's centroid
Ci: ith cluster
Some equations used in algorithm are:

$$|Ci,Cj| = \{d(mi,mj) : (i,j)\varepsilon\ [1,k] \ \&\ i \neq j\}$$

Where $|Ci,Cj|$ is the distance between cluster Ci and Cj.

$$dc(i) = \tfrac{1}{2}(\min\{|Ci,Cj|\})$$

where dc(i) is the half of the minimum distance from ith cluster to any other remaining cluster

iii. *Cluster Seed Selection*

When calculating the K turn of clustering seeds with the improved algorithm, those data in the cluster having a great similarity to the K-1 category seeds should be adopted to calculate their mean points (geometrical center) as the clustering seed of the K tum and the specific calculation method is below as[13]:

1) For the cluster Ci(k-1) obtained through the K-1 tum of clustering, the minimum similarity sim _ mini(K-1) of the data in the cluster to the clustering seed Si(k-l) of the cluster is calculated;
2) The data in the cluster Ci(k-l) is calculated that has a similarity of more than 1-β* (1-sim _ mini(k-l))to the clustering seed Si(k-l) (among, β is a constant between 0-1), and the data set is recorded as cni(k-1)
3) The mean points of the data in cni(k-l) are calculated as the clustering seed of the K tum.

b) *Methods to Define no of Cluster*

i. *Initialization Method*

This method depends on the data and works well to find the best number of cluster and their centroids values. It starts by reading the data as 2D matrix, and then calculates the mean of the first frame size F1=300x300,F2=150x150, F3=100x100, F4=50x50, F5=30x30, F6=10x10 or F7=5x5. Then, it keeps the value of means in an array called means array

even at the end of the data matrix. After that it sorts the values in the means array in an ascending manner. In cases where the values are similar, they are removed to avoid an overlap. In other words, only one value is kept. It will then calculate the number of elements in the means array: this number is the number of clusters and their values are the centroids values as indicated in the steps below as [11]:

1) Read the data set as a matrix.
2) Calculate the means of each frame depending on the frame size and putting them in the means array.
3) Sort the means array in an ascending way.
4) Comparing between the current element and the next element in the means array. If they are equal, then keep the current element and remove the next, otherwise, keep both.
5) Repeat step 4 until the end of the means array.
6) Count how many elements remain in the means array. These are equal to the number of clusters and their values.

ii. *The Encoding Method*

According to the characteristics of K-mean cluster algorithm, to find the optimum cluster, the optimum K value should be found, the value of K is the learning object of the genetic algorithm, the encoding it encoding to K value. In general situation, to the class issue, there is always a maximum number of classes "MAXCLassnum" for the cluster, this value is input by the user. So K is a integral between 1 and MAXCLassnum, can be indicated in a binary string. In this experiment, using a byte to express K value, that is 255 classes maximum. This value is enough for normal cluster problem[4].

1) Chooses n number of chromosome from the original n chromosome Using the roulette wheels selection of the traditional genetic algorithm.
2) Crossover method is applied on selected chromosome in the matting pool.
3) Mutation is applied over chromosomes in the matting pool.
4) Form a new generation of chromosome with the original chromosome.
5) Design the fitness function to evaluate K value by the quality of sample cluster result.

The Fitness function is:

$$Fitness = \omega_1 \frac{Dis\ \ of\ \ class}{1 + Dis\ \ in\ \ class} + \omega_2 \frac{1}{NumDifference}$$

Distance between classes is:

$$Dis\ \ of\ \ class = \frac{2\sum_{i=0}^{k}\sum_{j=i+1}^{k} dis(center_i, center_j)}{k(k-1)}$$

The centeri is classi of cluster center, dis(x,y)is the Euclidean distance between x,y.

Distance between classes is:

$$Dis\ \ in\ \ class = \frac{1}{k}\sum_{i=0}^{k}\left(\frac{\sum_{j=0}^{num_i} dis(Sample_{ij}, center_i)}{num_i}\right)$$

the numi is number of class of i, sample is the samplej of the classi, NumDifference show the statistics of the difference of sample between classes.The ω1, ω1 of Fitness function is the weight of distance between classes.

iii. *Tentative Clustering*

Clustering uses principal components analysis, to determine a tentative value of count of classes and provide changeable labels for objects. The kernel based clustering approach performs principal component analysis on standard score of a given matrix and thereafter projects the matrix into space of the calculated principal vectors. Count of the employed principal vector is depending on the given number of classes (K-Means algorithm needs 'K' to process). In order to avoid depending on the number of classes 'K' and to find maximum possible classes, we project the matrix to the space of all principal vectors. After we calculate a probability matrix (P) from result of projection (matrix C) such that Pi,j entry shows probability of connectivity of ith object to jth object.

By refining matrix C according to the probability values of matrix P, we find a block matrix that represents groups of objects[12].

V. Conclusion

This paper presents an overview of the k-means clustering algorithm. K-means clustering is a common way to define classes of jobs within a Dataset. The initial starting point selection may have a significant effect on the results of the algorithm, both in the number of clusters found and their centroids. Methods to improve performance of k-means clustering are discussed in this paper. These methods fall into two categories: initial point selection and define number of cluster. Six of these methods, three from each category, are presented. These methods have been implemented in data mining system and can get better results for some practical programs such as character recognition, image processing, text searching .

References Références Referencias

1. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
2. Jirong Gu; Jieming Zhou; Xianwei Chen;"An Enhancement of K-means Clustering Algorithm", in proceeding of Business Intelligence and Financial Engineering, 2009. BIFE '09. International Conference ,2009, Key Lab. of the Southwestern Land Resources Monitoring, Sichuan Normal Univ., Chengdu, China

3. Singh, R.V.; Bhatia, M.P.S.;" Data Clustering with Modified K-means Algorithm"in proceeding of Recent Trends in Information Technology (ICRTIT), 2011 International Conference,2011, Dept. of Comput. Sci. & Eng., Univ. of Delhi, New Delhi, India
4. Juntao Wang; Xiaolong Su; "An improved K-means Clustering Algorithm", in proceeding of Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference,2011, Sch. of Comput. Sci. & Technol., China Univ. of Min. & Technol., Xuzhou, China
5. Colm Kearney and Andrew J. Patton, "information gain ranking", Financial Review, 41, 2000, 29-48
6. D. Mła yszko, S. T. Wierzchoń "Standard and Genetic K-means Clustering Techniques in Image Segmentation", (CISIM'07) 0-7695-2894-5/07 IEEE 2007
7. C. M. Bishop, " Neural networks for pattern recognition ", Clarendon Press, Oxford, 1995.
8. B. Zhang, "Generalized k-harmonic means - Boosting in unsupervised learning", Technical Report HLP-2000-137", Hewlett-Packard Labs,2000.
9. J. Pérez, R. Pazos , L. Cruz, G. Reyes, R. Basave, and H. Fraire "Improving the fficiency and Efficacy of the K-means Clustering Algorithm Through a New Convergence Condition", Gervasi and M.Gavrilova (Eds.): ICCSA 2007, LNCS 4707, Part III, pp. 674–682. Springer-Verlag Berlin Heidelberg 2007.
10. W. Li "Modified K-means clustering algorithm", 978-0-7695-3119-9/08 , 2008 IEEE, DOI 10.1109/CISP.2008.349
11. Samma, Ali Salem Bin; Salam, Rosalina Abdul;" daptation of K-Means Algorithm for Image Segmentation"in proceeding of International Journal of Signal Processing,2009
12. Dashti, H.T.; Simas, T.; Ribeiro, R.A.; Assadi, A.; Moitinho, A.; "MK-means - Modified K-means clustering algorithm", in proceeding of Neural Networks (IJCNN), The 2010 International Joint Conference,2010, Univ. of Wisconsin, Madison, WI, USA
13. Li Xinwu;" Research on Text Clustering Algorithm Based on Improved K-means",in proceeding of Computer Design and Applications (ICCDA), 2010 International Conference ,2010, Jiangxi Univ. of Finance & Econ., Nanchang, China

This page is intentionally left blank